Machine-Learned Interatomic Potentials for Photophysics

Lucas Sánchez Garay

Department of Physics, University of Warwick, Coventry CV4 7AL, United Kingdom

13/03/2025

This study focuses on fine-tuning the MACE-OFF23 model, a foundational Graph Neural Network (GNN) initially trained on ground-state Potential Energy Surfaces (PES), to accurately predict excited-state PES for photophysical processes. We apply this fine-tuned model to two systems: the oxirane angle opening reaction and thymine dimerisation, both of which involve ultrafast photoinduced excitations. We address challenges such as accurate prediction of excited-state dynamics. Our results demonstrate that fine-tuning foundational models with a small dataset can achieve accuracy comparable to models trained from scratch on a large dataset, while also highlighting the limitations in predicting internal conversions and conical intersections. This work underscores the potential of transfer learning and multi-head models in simulating complex photochemical reactions.

I. Introduction

Machine-Learned Interatomic Potentials have recently emerged as a more efficient method to perform quantum predictions whilst maintaining accuracy. Calculations on forces and energies in molecules are normally performed using Density Functional Theory. This method, despite being highly accurate, remains very computationally expensive. The machine learning alternative is much faster, computationally cheaper and can be used to predict optical spectra, transition states, and decay pathways among many applications.

A recent deep learning model, MACE-OFF23, has been developed as a "foundational model". This allows the model to be re-trained to make predictions in areas beyond its original training scope. The foundational model was initially trained on large datasets of QM predictions and is exceptionally good at predicting ground-state Potential Energy Surfaces [1].

The aim of this project is to fine-tune MACE-OFF23 for accurate predictions on excited-state Potential Energy Surfaces. The purpose of this paper is ultimately to perform excited molecular dynamics on two systems, oxirane angle opening reaction, and thymine dimerisation.

II. Theoretical Background

A. Photophysics and Ultrafast Processes

Photochemical reactions involve light-induced molecular transformations that excite molecules to higher-energy states. These reactions occur on an ultrafast timescale, typically within femtoseconds. Understanding these rapid processes is critical for the advancement of technologies such as optical memory systems, where precise control of molecular photoswitching materials is essential [2].

Data obtained by experiments measuring photochemical processes are difficult to interpret and lack the theoretical modelling needed to fully understand them. Hence, the development of computational methods to simulate these reactions is the key to understanding excited-state dynamics [3].

The theory developed so far stems from the need to solve Schrödinger's equation for many-atom systems. This was aided by the Born-Oppenheimer approximation, by which it is possible to separate the motion of the nuclei from the electrons, decoupling the electronic and nuclear degrees of freedom [4]. This approximation enables the replacement of the quantum mechanical wavefunction for the nuclei with a classical picture.

However, this approximation breaks down in regions where energy states are degenerate, regions where states cross paths. These regions became known as conical intersections and proved to be highly relevant in the radiationless decay of a photoexcited state to a stable ground state, among other cases [5].

This is of particular relevance in molecular biology, as DNA bases are reactive to UV light and are harmed by it, such as in thymine dimerisation. However, DNA bases have also been shown to be photo stable [6], as a result of efficient radiationless decay channels into stable ground states, which will be delved on further into this paper.

The computational study of molecular processes have given rise to a new theoretical branch, Density Functional Theory.

B. Density Functional Theory

Density Functional Theory, a type of *ab initio* models, is the theory by which quantum predictions on a given system are performed bottom-up, in other words, by performing calculations on electron density rather than on wave functions. At its foundation lies the Many-Body Schrödinger Equation, which, under the Born-Oppenheimer approximation, can be expressed as:

$$\hat{H} = \sum_i \frac{\mathbf{p}_i^2}{2m_e} + \sum_{i>j} \frac{e^2}{4\pi\epsilon_0|\mathbf{r}_i - \mathbf{r}_j|} - \sum_i \sum_\alpha \frac{Z_\alpha e^2}{4\pi\epsilon_0|\mathbf{r}_i - \mathbf{R}_\alpha|} + \sum_{\alpha>\beta} \frac{Z_\alpha Z_\beta e^2}{4\pi\epsilon_0|\mathbf{R}_\alpha - \mathbf{R}_\beta|} \tag{1}$$

The first summation represents the kinetic energy of the electrons, where $\mathbf{p}_i$ denotes the momentum of each electron in the system. The second summation corresponds to electron-electron interactions, with $\mathbf{r}_i$ denoting the position of each electron. The third summation accounts for nucleus-electron interactions, where Z_α and $\mathbf{R}_\alpha$ represent the atomic number and position of nucleus α, respectively. The final summation describes nucleus-nucleus interactions.

However, solving this Hamiltonian is computationally intensive. Therefore, the fundamental approach shifts from solving for a wave function to solving for the electronic density [7, 8]. Utilizing the common quantum mechanic generalization of electron density as:

$$n(\mathbf{r}) = \int \psi^*(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_n) \left[\sum_{i=1}^{N} \delta(\mathbf{r} - \mathbf{r}_i) \right] \psi(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_n) d\mathbf{r}_1, ..d\mathbf{r}_n = \sum_j |\psi_j(\mathbf{r})|^2 \tag{2}$$

In 1964 Hohemberg and Kohn formulated two theorems that established the foundational principles of Density Functional Theory. The first theorem states that the ground state energy of a many-electron system is uniquely dependent on electron density $\rho(\mathbf{r})$. Implying there is a one-to-one mapping between the ground-state density and ground state wave function [9, 10].

The second theorem states that there is a universal energy functional of the electron density $E[\rho]$. The result of this functional gives the unique ground state energy of the system [10].

Kohn-Sham later developed a set of equations to solve the Many-Body Schrödinger Equation based on these theorems [11]:

$$\left[-\frac{1}{2}\nabla^2 + V_{\text{ext}}(\mathbf{r}) + V_{\text{H}}([n];\mathbf{r}) + V_{\text{xc}}([n];\mathbf{r})\right]\psi_k(\mathbf{r}) = \epsilon_k\psi_k(\mathbf{r}) \tag{3}$$

Where T is the kinetic energy, V_{ext} is the external potential, V_{H} is the Hartree potential that accounts for particle-particle or Coulomb interactions, and V_{xc} is the exchange-correlation potential to account for other interactions as well as the antisymmetric nature of a wave function for identical fermions [8].

Kohn-Sham equations are simpler and more efficient to solve. The only remaining unknown term is the exchange-correlation functional V_{xc}, whose exact form is unknown. Although numerous approximations and various functionals have been developed for different tasks, all calculations in this paper have employed the same exchange-correlation functional (xc) and basis-set: a hybrid van der Waals density functional, ωB97M-D3(BJ) [12]; the basis set employed is the diffuse basis set def2-TZVPPD [13]. This basis and functional combination was deliberately chosen to match those used by in the initial training of the foundational model [1].

To study excited-state properties of systems, DFT alone is insufficient. Although it excels in ground-state calculations, it does not provide accurate descriptions of excited states. Advanced methods such as time-dependent DFT (TDDFT) are needed for reliable predictions. In this paper we use 2 different TDDFT theories to finetune the foundational model on excited states, Conventional Linear Response TDDFT and Spin-Flip Linear Response TDDFT.

B..1 Time Dependent Density Functional Theory (TDDFT)

Time Dependent Density Functional Theory builds upon ground-state DFT by extending calculations to time-dependent systems. It utilises Perturbation Theory to analyse a system's response when subject to a small external field perturbation. The basis of TDDFT lies in the Runge-Gross theorem, which serves as an equivalent to the Hohenberg-Kohn theorems for time-dependent systems [14]. The time dependent Kohn-Sham equation is easily obtainable by rewritting equation 3 to account for time:

$$\left[-\frac{1}{2}\nabla^2 + V_{\text{ext}}(\mathbf{r},t) + V_{\text{H}}([n];\mathbf{r},t) + V_{\text{xc}}([n];\mathbf{r},t)\right]\psi_k(\mathbf{r},t) = i\frac{\partial}{\partial t}\psi_k(\mathbf{r},t) \tag{4}$$

Where ψ_k denotes the individual Kohn-Sham orbitals. This gives rise to a new definition of time dependent electron density:

$$\sum_k |\psi_k(\mathbf{r}, t)|^2 = n(\mathbf{r}, t) \tag{5}$$

The solution to the time-dependent Kohn-Sham equations 4 5 requires self-consistency, as the potentials V_H and V_{xc} depend on the electron density, which is obtained from the sum of the Kohn-Sham orbitals in Equation 5. This equation is solved iteratively by first selecting an initial set of orbitals $\left\{\psi_k^{(t)}\right\}$, using them to compute the electron density $n^{(t)}$, and subsequently updating the orbitals to $\left\{\psi_k^{(t+1)}\right\}$ and the electron density to $n^{(t+1)}$. This process continues until convergence is achieved, when $\left\{\psi_k^{(t')}\right\}$ and $\left\{\psi_k^{(t'+1)}\right\}$ differ by less than a predefined tolerance [14, 15].

However, in the time domain, the exchange-correlation potential is a complicated functional of the electron density n, dependence on space and time. There are a number of approximations for solving this equation. However, as mentioned above, we will only focus on two: Conventional Linear Response and Spin-Flip Linear Response.

B..2 Conventional Linear Response TDDFT

The conventional linear response approach is to define an external potential acting as a perturbation which can then approximate the electron density as a Taylor expansion. We can define a potential as [14]:

$$v_{\text{ext}}(\mathbf{r}, t) = v_{\text{ext}}(\mathbf{r}, t = 0) + \delta v_{\text{ext}}(\mathbf{r}, t) \tag{6}$$

where $v_{\text{ext}}(\mathbf{r}, t = 0)$ corresponds to the external potential at $t = 0$ and $\delta v_{\text{ext}}(\mathbf{r}, t)$ represents the perturbation. The Taylor-series approximation response of electron density is then:

$$n(\mathbf{r}, t) = n_{\text{GS}}(\mathbf{r}, t) + n_1(\mathbf{r}, t) + ... \tag{7}$$

Linear Response Theory is concerned with the first-term $n_1(\mathbf{r}, t)$ and n_1 is computed from density-density linear response function χ as:

$$n_1(\mathbf{r}, t) = \int_0^\infty dt' \int d^3r' \chi(\mathbf{r}t, \mathbf{r}'t') \delta v_{\text{ext}}(\mathbf{r}', t'), \quad \text{where} \quad \chi(\mathbf{r}t, \mathbf{r}'t') = \left.\frac{\delta n(\mathbf{r}, t)}{\delta v_{\text{ext}}(\mathbf{r}', t')}\right|_{v_{\text{ext}},0} \tag{8}$$

Without going through the full derivation, equation 8 can be converted into Cassida's equation to calculate excitation energies. Cassida's equation is expressed as [8]:

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} = \omega \begin{pmatrix} \mathbf{X} \\ -\mathbf{Y} \end{pmatrix}. \tag{9}$$

where ω is the excitation energy eigenvalue we are solving for, $\mathbf{A}, \mathbf{B}$ are matrices involving derivatives of the Fock matrix, and $\mathbf{X}, \mathbf{Y}$ are vectors containing transition amplitudes. The detailed derivation of these terms will not be discussed here. The Tamm-Dancoff Approximation (TDA) is commonly used in Linear Response TDDFT setting $\mathbf{B}$ to zero.

However, linear-response theory has limitations near conical intersections, where energy states become degenerate, and it cannot accurately describe double excitations. To solve these limitations, the second theory Spin-Flip will be used.

B..3 Spin-Flip Linear Response TDDFT (SF-TDDFT)

Spin-Flip Linear Response TDDFT is based on the concept of using a high-spin reference state, where the total spin quantum number of the reference state S is one unit higher than that of the target state of interest $S - 1$. The final target states are calculated through the Spin-Flipping excitations, $\alpha \rightarrow \beta$ from the reference state [16].

Spin flip calculates excitation energies using the same formalism as Linear Response Theory; employing Casida's equation (9). However, it incorporates Spin-Flip transitions, where an electron is excited from a spin-up orbital (α) to a spin-down orbital (β). This Spin-Flip mechanism allows the method to capture transitions that involve changes in spin multiplicity and total spin, such as singlet-triplet or triplet-singlet excitations. These Spin-Flip transitions are crucial in cases of conical intersections, where changes in spin multiplicity can occur, such as singlet-triplet or triplet-singlet excitations. In contrast, linear response theory is limited in this respect, as it does not account for spin-flipping processes. Consequently, Spin-Flip TDDFT is better equipped to resolve electronic states near conical intersections, where spin changes are an important feature [16].

One potential drawback of Spin-Flip methods is the risk of spin contamination, where the wavefunction includes contributions from incorrect spin multiplicities, leading to inaccuracies. In such cases, a mixed reference Spin-Flip approach can help by combining a higher spin reference state with a second lower spin reference state, reducing contamination [17]. While SF-TDDFT addresses the topology problem around conical intersections

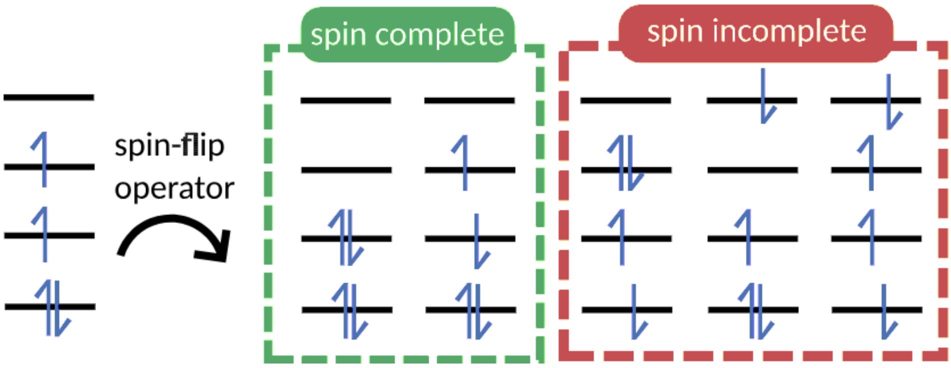


Figure 1: Effect of Spin-Flip on a $S = 3$ system [16]

that conventional Linear Response TDDFT faces, it significantly worsens the spin contamination issue. This occurs because some double excitations are admitted to the excitation space in a manner that excludes the necessary determinants to form pure $\hat{S}^2$ eigenstates [18]. This was found to be the case for a thymine dimer excited state, which will be discussed further.

Despite these advancements, solving Casida's equation for both linear response and Spin-Flip scenarios remains computationally demanding for larger or more complex systems. This highlights the need for a different technique that can provide efficient and accurate approximations of excitation energies without the computational burden of solving such equations directly. Such techniques are based on deep-learned models trained on large datasets to predict PES without directly solving Casida's equation.

In order to train a model on the PES of a system, realistic configurations of the system need to be sampled. This is done by running geometry optimisations to relax a configuration given a constraint. Constraints are used to ensure the configurations sampled do not all correspond to the Frank-Condon point, the equilibrium geometry of the initial electronic state [19]. The constraints used in thymine dimer and oxirane will be addressed later in the report.

B..4 Geometry Optimisations

When exploring the subspace of a molecular system around given coordinates, constraining a number of variables in the system it is necessary to ensure the system is in its relaxed lowest energy state. As these represent stable molecular structures where forces on the atoms are minimised, ensuring that the potential energy surface is well-defined and physically meaningful. This implies finding the positions of the atoms in the system that have the lowest potential energy. To obtain such geometries, geometry optimisations

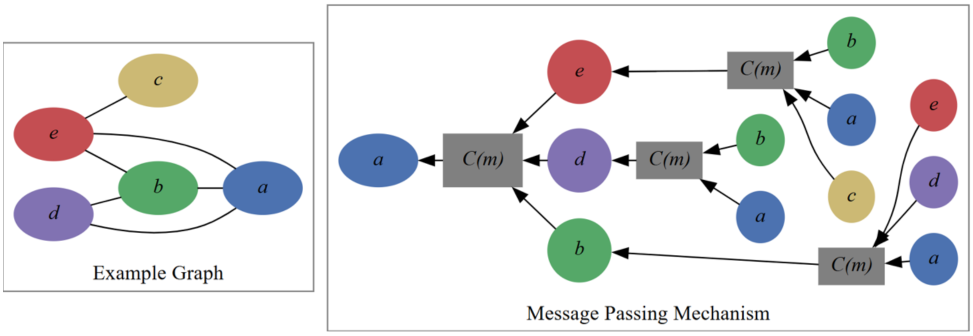


Figure 2: Message-passing mechanism illustrated in a sample network, with node a as the target node and $C(m)$ a placeholder for the message aggregation [24].

have been employed. These are performed by using numerical optimisation algorithms, to find the lowest energy positions of the individual atoms of the system while respecting a constraint. In this case the algorithm employed is BFGS [20].

C. Machine-Learned Potentials, GNN's and MACE-OFF23

The use of machine learning in calculating PES began in the 1990s, marked by the publication of the first Machine-Learned Potential (MLP) by Doren et al. [21]. Their pioneering model used a simple feed-forward neural network to approximate PES. The paper showed promising results in the capture of short-range molecular interactions. Early MLPs achieved significant accuracy for localized interactions [22], but faced limitations in predicting non-local and long-range interactions, which are critical for describing phenomena such as dispersion forces.

These shortcomings prompted further development of more sophisticated machine learning architectures. Newer models employ an architecture named Graph Neural Networks (GNN), as the final goal in all models is to map a graph of different chemical elements to a potential energy [23].

GNNs work by representing a molecule as a graph, where the nodes are the atoms of the molecule. Atoms lying at a distance smaller than r_{cutoff} are set to be neighbours and connected by an edge [23].

GNNs infer and predict properties using a mechanism called message passing. Message passing works by using connections in the graph to understand interactions of nodes. It begins with each node in the graph sending a message to its neighbours, who then aggregate all the messages and send a new message to its neighbours. This process is repeated with each layer of the GNN to make higher-level predictions, figure 2 illustrates the mechanism.

A more sophisticated GNN architecture is the one used in MACEOFF, which is based on a smarter technique involving symmetries of the system called equivariant message passing.

C..1 Equivariant Message Passing and MACE-OFF23

The foundational model that we will fine-tune for excited-state predictions is MACE-OFF23 [1]. Developed at the start of 2024, MACEOFF is a GNN that utilises a technique called equivariant message passing, an extension of traditional message passing that incorporates the symmetries inherent in the system. This method leverages embeddings at each node to encode and communicate higher-order information by constructing messages that respect the system's symmetry properties. Specifically, MACE-OFF23 employs spherical harmonics and symmetry conservation principles to generate and interpret these messages effectively. This is extremely important as different predicting classes behave differently under rotation or translation actions. Given an arbitrary system, if we rotate the system, its magnetic dipoles will change while its internal energy will stay the same. The state of a node (atom), in the network, is represented by the tuple:

$$\sigma_i^{(t)} = (\boldsymbol{r}_i, z_i, \boldsymbol{h}_i^{(t)}) \tag{10}$$

Where the superscript (t) represents the layer index of the network, $\mathbf{r}_i$ is the position of the atom, z_i the chemical element and $\mathbf{h}_i$ the learnable features. the node features are initialised as a learnable embedding of the chemical elements z_i into the hyperparameter number of channels k:

$$\boldsymbol{h}_{i,k00}^{(0)} = \sum_z W_{kz} \delta_{zz_i} \tag{11}$$

where the subscript 00 represent the spherical harmonics indices l, m, initially set to 0. W_{kz} represents the matrix that maps the atomic number z_i to the initial node features. An issue arises with this technique because for systems with many unique atomic numbers the matrix W_{kz} grows and the embedding on each atom grows, making it computationally expensive for very large systems. A forward pass in the network involves 3 specific type of steps, *message construction*, *update* and *readout*. The *message construction* is performed as follows:

$$\boldsymbol{m}_i = \bigoplus_{j \in \mathcal{N}(i)} M_t(\sigma_i^{(t)}, \sigma_j^{(t)}) \tag{12}$$

Where $\oplus$ is a learnable, permutation invariant pooling over the neighbouring atoms of i $\mathcal{M}(i)$ and M_t is a learnable message function. In the *update* step the learnable features of a node $\boldsymbol{h}_i^{(t)}$ are updated via:

$$\boldsymbol{h}_i^{(t+1)} = U_t(\sigma_i^{(t)}, \boldsymbol{m}_i^{(t)}) \tag{13}$$

Where U_t is a learnable update function. The *message passing* and *update* steps utilise a one particle basis that is formed from the combination of Clebsch-Gordan coefficients and spherical harmonics to ensure equivariance is preserved. The full explanation on the one particle basis, message construction and update are beyond the scope of this project and can be found in [1].

Once the network has forward-passed through all the layers the *readout* step maps the node states to the target. Since MACE-OFF23 only has 2 layers the readout function is defined as a linear combination of rotationally-invariant node features of the first layer and a multi-layer perceptron for the second layer. Hence the energy of a system is read as:

$$E_{tot} = \sum_i E_i = \sum_i \sum_{t=1}^{2} E_i^{(t)} = \sum_i \sum_{t=1}^{2} \mathcal{R}^{(t)}\left(\boldsymbol{h}_i^{(t)}\right) \tag{14}$$

With the *readout* function $\mathcal{R}^{(t)}$ defined as:

$$\mathcal{R}^{(t)}\left(\boldsymbol{h}_i^{(t)}\right) = \begin{cases} \sum_k W_k^{(t)} h_{i,k00}^{(t)} & \text{for } t = 1 \\ \text{MLP}\left(\left\{h_{i,k00}^{(t)}\right\}_k\right) & \text{for } t = 2 \end{cases} \tag{15}$$

The forces on the atoms are obtained by analytically calculating the derivatives of the potential energy with respect to the position.

This architecture is what makes MACEOFF excel at organic chemistry quantum predictions, as the higher order and equivariance preservation allows messages to contain more insightful information into the behaviour of a molecule.

Although MACE-OFF23 excels at efficient and cost-effective calculations of ground-state properties with DFT-level accuracy, it has not been trained on excited states, geometries with energy degeneracies, or conical intersections. Our work aims to address this limitation by fine-tuning the model for these specific scenarios.

III. Methods

A. Data Curation

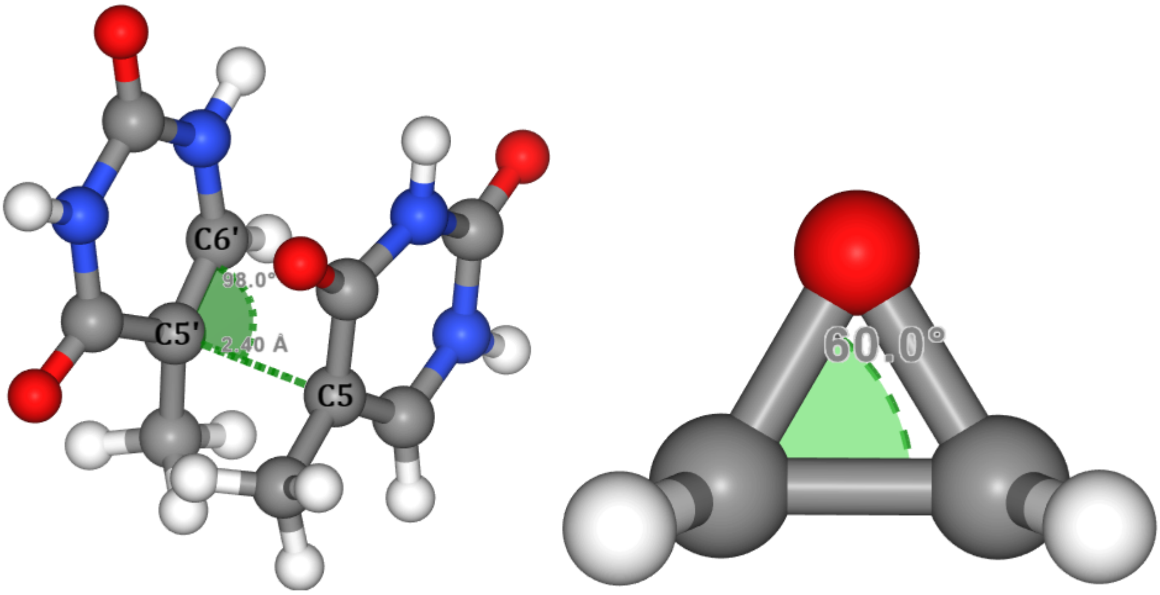
The data to be used to train the model were obtained by performing Spin-Flip TDDFT calculations on a number of configurations. All TDDFT-related calculations in this paper were performed in the ORCA program system [25]. The TDDFT calculations are performed with the *%tddft* module. In said module you must specify the root number you want to solve for, this will specify the excited state ORCA is calculating energies for. A problem occurred while solving for $I_{\mathrm{root}} = 0$, which corresponds to the ground state in Conventional Linear Response and the first excited triplet state in Spin-Flip Linear Response. Specifically, the forces on the atoms within the system were not computed and erroneously appeared as zero. This was solved by calculating ground state energies and forces using common DFT and not using the TDDFT module for Conventional Linear Response and using $I_{\mathrm{root}} = 1, 2$ for Spin-Flip Linear Response, which corresponds to the singlet ground state and singlet first excited state respectively [25].

However the version of ORCA used throughout the project, version 6.0.1, lacks an implementation of Mixed Reference Spin-Flip TDDFT which would have allowed for pure excited states in the case of dimersied thymine.

The initial plan was to perform two types of TDDFT calculations, conventional Liner Response TDDFT and Spin-Flip Linear Response TDDFT. However it was found that training the model on the two levels of theory discussed to predict ground and first excited states limited the model's ability to predict accurate energies on both levels. As a result, only Spin-Flip Linear Response TDDFT would be used.

The configuration space explored was purposely selected to be around a conical intersection in an attempt to teach the model how to handle molecular geometries in which the ground-state energy is at the same energy level or higher than the first excited state. A key aspect of data curation was ensuring that the geometry optimisations converged to physically reasonable structures, avoiding cases where the molecule was fragmented or where atoms were positioned unrealistically close together, leading to excessively high potential energies.

To ensure that geometries were optimised for realistic configurations, we compared the potential energy of the molecule as a function of the constraint coordinates with existing literature.



(a) Thymine Dimer

(b) Oxirane

Figure 3: Left Panel shows the geometry of di-thymine showing the C5-C5' bond length (2.4Å) and the C5-C5'-C6' angle (98°). Coordinates used to explore the Potential Energy Surface and perform the geometry optimisations. Right panel shows the geometry of oxirane. Angle showed is the angle to be opened during optimisations.

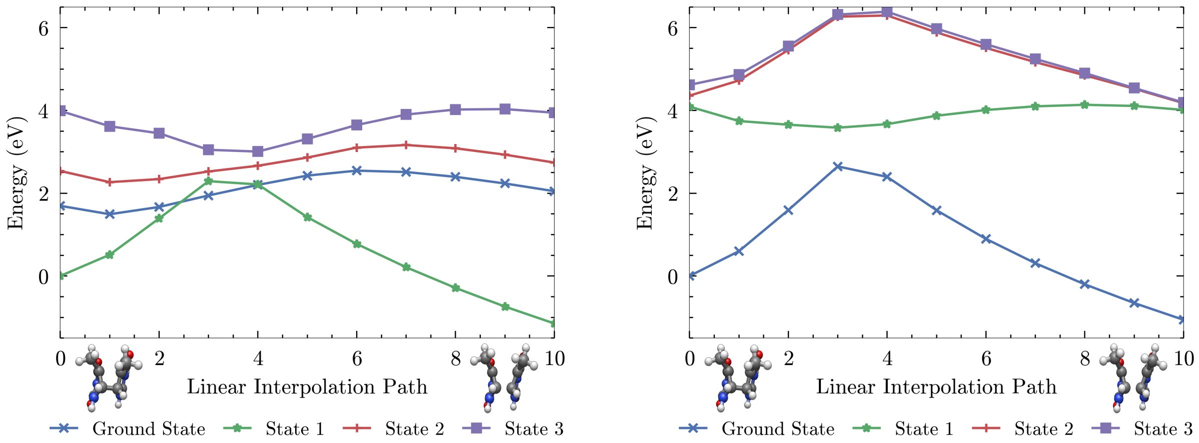
A..1 Thymine Dimer

For the thymine dimer system, the configuration space explored involved geometry optimisations along two key coordinates: the bond length of the covalent C5-C5' bond between the two thymine molecules and the angle formed by the bonding carbons (C5-C5') and one of the adjacent carbons (C5-C5'-C6') relative to the bonding carbons. These two coordinates are illustrated in figure 3a. In the following sections of the paper, the geometries of thymine dimer will be denoted as di-Thy(C5-C5', C5-C5'-C6'). For the specific geometry shown in the figure, this would be referred to as di-Thy(2.4, 98).

The subsample space consisted of a combination of 12 different C5-C5' bond lengths and 8 different C5-C5'-C6' bond angles, giving a total of 96 unique configurations for thymine dimer. In that space there are multiple S0/S1 crossings, for simplicity consider the linear interpolation between geometries di-Thy(2.2, 76) and di-Thy(2.6, 76).

Figure 4 shows the linear interpolation along the splinting of the C5-C5' bond. Furthermore, the sub figure 4a shows the conical intersection along the path which is in agreement with the results obtained in [26].

During the data curation process for the first excited state of di-thy, a diverse set of configurations was identified, exhibiting different spin multiplicities. A subset of these



(a) Spin-Flip

(b) Linear Response

Figure 4: Linear Interpolation Path calculations between di-Thy(2.2, 74) and di-Thy(2.6, 74).

configurations showed varying degrees of spin purity, including cases corresponding to the singlet first excited state (S_1) and the triplet first excited state (T_1). This arises due to the unavailability of mixed-reference Spin-Flip methods in ORCA, which constrains the ability to fully control spin contamination in certain cases.

A..2 Oxirane

For oxirane, the coordinate space explored was along the angle opening reaction, by opening the C-C-O angle until the C-O bond opposite to the opening angle breaks. Figure 3b shows the angle to be opened during geometry optimisations

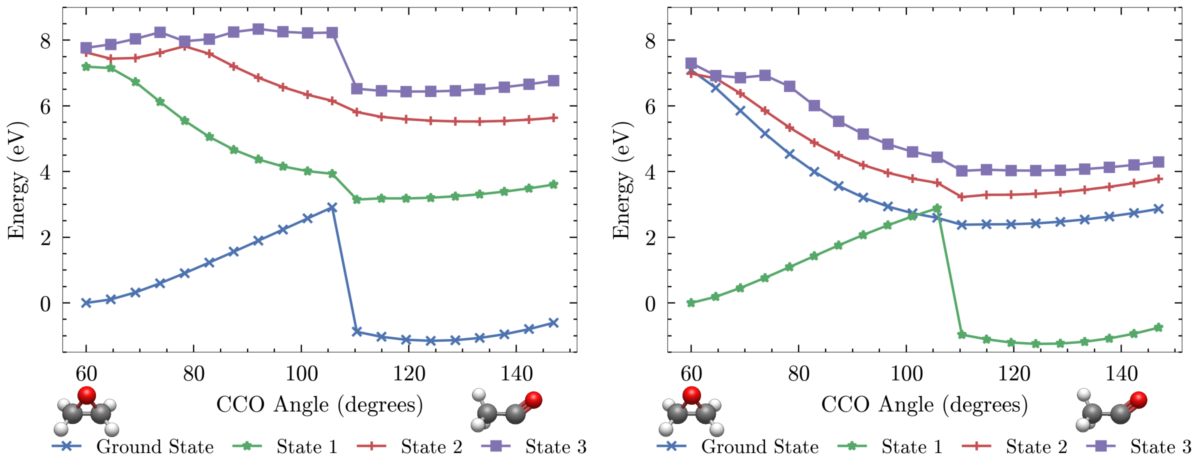
Figure 5 shows the PES of the opening of the angle. Comparing the two graphs shows how Conventional Linear Response lacks the conical intersection present of Spin-Flip.

Additionally, figure 5 illustrates the difference between Conventional Linear Response and Spin-Flip Linear Response. The Spin-Flip method provides better resolution near the conical intersection at 108°. In the Spin-Flip panel, State 1 appears to correspond to the ground state of Conventional Linear Response due to the higher reference state mentioned in section B..3.

The results shown in Figure 5 are consistent with [27], which confirms that the potential energy surface of oxirane is correctly mapped.

B. Training

Once the datasets for thymine dimer and oxirane were curated, the first fine-tuned models of MACE-OFF were ready for training. However, during the initial training phase, issues



(a) Linear Response

(b) Spin-Flip

Figure 5: Oxirane constrained optimisation scan for varying CCO angles, followed by TDDFT calculations: Left panel shows Conventional Linear Response results, right panel displays Spin-Flip Linear Response results. The figure highlights differences between SF and LR methods in predicting states near a conical intersection.

arose, primarily due to a common challenge encountered when fine-tuning foundational models: catastrophic forgetting. When new training data is introduced to a pre-trained model, the model tends to forget the previously learned data. This results in a loss of knowledge, making it difficult for the model to retain information from earlier tasks while learning from the new dataset.

The MACE environment provides a solution to this issue called replay fine-tuning. This method involves adding an additional output head trained on a subsample of the original training data to preserve the knowledge acquired during the initial training. However, given the recency of MACE-OFF, replay fine-tuning had only been developed for the MACE-MP family of models and needed to be adapted for MACE-OFF.

The implementation of replay fine-tuning involved several key steps. First, the initial training data was filtered to include only configurations containing the same molecular species as those in the new fine-tuning dataset. For thymine dimer, this meant selecting configurations with C, O, H, and N, while for oxirane, configurations with C, O, and H were chosen. The next step involved balancing the subsampling of the original training data. A sufficient number of configurations were selected to preserve previously learnt knowledge while keeping the dataset size relatively small to avoid excessively long training times.

As a result, the fine-tuned models developed have a structure consisting of three output heads: one trained on ground-state Spin-Flip TDDFT, another trained on first-excited-state Spin-Flip TDDFT, and a final head trained on a subsampling of the original data. Subsampling was performed using the data set found in [28].

The hyperparameters used for fine-tuning the thymine dimer and oxirane models were selected based on the recommendations of [1] and the MACE library documentation. The specific values are listed in Table 1.

Hyperparameter	Value
Number of epochs	700
Batch size	5
Epoch start SWA	475
r_{max}	5.0

Table 1: Hyperparameters used during training *Epoch start SWA* refers to the epoch at which the energy weight of the loss is increased by $100x$, r_{max} controls the cutoff radius, atoms separated by more than r_{max} are not seen as neighbours.

C. Molecular Dynamics and Active Learning

Until now, all the data given to the model for training has been on relaxed geometries, whilst this is essential to ensure that the model can accurately describe the potential energy surface of a system, meaning it might not be able to handle systems with a higher energy than its relaxed state.

To avoid this shortcoming, an active learning mechanism was implemented to teach the model how to deal with energetic configurations. In this process, a molecular dynamics simulation was set were a configuration close to the conical intersection is selected. The fine-tuned model trained on relaxed geometries runs a molecular dynamic simulation with a step size of 0.5 fs ($5 * 10^{-16}$ s) for 500 fs, snapshots of the positions where taken at 5, 10, 50, 200 and 500 fs.

The TDDFT energies of the ground and first excited state were calculated and compared with the predictions of the model. If the model energy predictions on a given snapshot differed by more than 25% compared with the TDDFT energies, the snapshot was saved to retrain the model.

This process was performed for 10 simulations, given 5 snapshots per configuration, and hence 50 configurations were used in fine-tuning. The initial configurations were selcted from the same geometry but added a Gaussian noise rattling to the atom positions to avoid data repetition.

Adding more energetic configurations to the training data allows the model to gain knowledge on predicting properties of systems that are not relaxed; this also has a drawback, the training error is higher when training on both relaxed and energetic configurations than when training models on the separated datasets. This is particularly the case in the first excited state, as the surface potential energy of the excited state is highly dependent on the internal energy of the system [29].

The combined dataset of relaxed and energetic configurations for each model, dimerised-thymine and oxirane is shown in table 2.

Model	Relaxed Samples	MD Samples	SPE	**Combined**
Thymine Dimer	96	50	4	**150**
Oxirane	100	50	3	**153**

Table 2: Combined dataset details for active learning, SPE refers to *Single Point Energies* are single atom reference energies, the kinetic energy of the atoms at $T = 0$ K [19]. 4 for dimerised Thymine, C, O, H, N: 3 for Oxirane C, O, H.

C..1 Multihead Molecular Dynamics

Another important aspect of the models under study is their capability to perform molecular dynamics simulations on both heads. Initially, the simulation is set to the first excited-state head, State 2. However, if during the simulation the ground state head, State 1, is found to have a higher energy than State 2, the simulation is adjusted to reflect the output from the State 1 head. This process is applied to all four models for a total simulation time of 500 fs, with a step size of 0.5 fs, at a temperature of 500 K.

D. Benchmarking

The benchmarking was performed by comparing 4 models for each system: (i) a fine-tuned model trained on relaxed & MD samples, (ii) a fine-tuned model trained on only relaxed samples, (iii) a fine-tuned model trained on only MD samples and finally, (iv) a scratch model trained on the combined relaxed & MD samples dataset. The last model is used to determine whether the knowledge of the foundational model proves useful when simulating excited molecular dynamics and as a benchmark of the maximum accuracy possible on the combined dataset.

In order to benchmark the 4 models 2 different techniques where used. The first one is similar to the one used in the active learning section, section III.C.. Where the fine-tuned model trained on the combined dataset was used to generate 250 configurations. The configurations were sampled as follows, a relaxed geometry close to a conical intersection was selected, a small Gaussian was then applied to the atom positions, Langevin molecular dynamics were ran for 500 fs and gemoetries were sampled at $5, 10, 50, 200, 500$ fs respectively.

For the resulting 250 configurations, ground state and excited state Spin-Flip TTDFT energies were calculated to be used as ground truths. For each model, the RMSE of each of the two heads was calculated using the ground truths.

The second stage of benchmarking involved examining how each of the models behaved when told to run molecular dynamics of a configuration close to a conical intersection at high temperatures. The simulations are set to high temperature as to ensure that the system has enough energy to overcome potential wells such that the molecular dynamics should not oscillate and a crossing of the ground and excited state potential energy surfaces is forced.

IV. Results

A. Thymine Dimer

A..1 Excited State Predictions

During the training of the thymine dimer models, it became evident that the model was not predicting a pure singlet excited state (S_1) but rather a mixture of singlet (S_1) and triplet (T_1) states. This spin contamination arises from the presence of double-excitations in thymine [30], which are not fully accounted for in the excitation space of Spin-Flip TDDFT. Consequently, the SF-TDDFT method admits certain double excitations that can lead to significant spin contamination because it does not include all the necessary determinants to form spin-pure states [18].

Two MACE-OFF23 models were trained on the same datasets one model being fine-tuned and another trained from scratch. Both models predict on the ground and first excited state of figure 4. Note that the predictions are made on the first and second excited states of the Spin-Flip reference states, which as mentioned in section B..3 accounts for the ground state of the singlet and the excited state of the first singlet, respectively.

Figure 6 shows the predictions of both models on the interpolation path geometries. The fine-tuned model is better at approximating both PES. However both models struggle to approximate image 3 of the interpolation, where the conical intersection is situated.

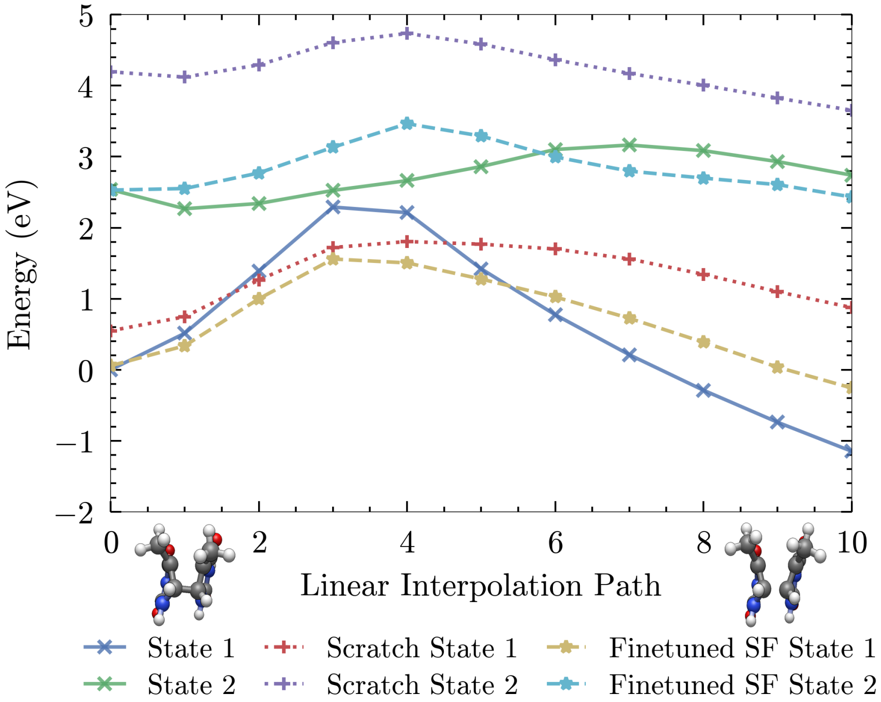


Figure 6: Scrath MACE-OFF vs Fine-tuned MACE-OFF predictions on the Linear Interpolation path of di-thy(2.2,74) and di-thy(2.6,74).

A..2 Multihead Molecular Dynamics

The thymine dimer molecular dynamics simulation of each model can be seen in figure 7. Given the significant kinetic energy of the simulations, as the temperature was set to 500 K the dynamics are expected to oscillate rapidly, this is not the case for the *Scratch model*, figure 7d showing the effectiveness of foundational model transfer learning.

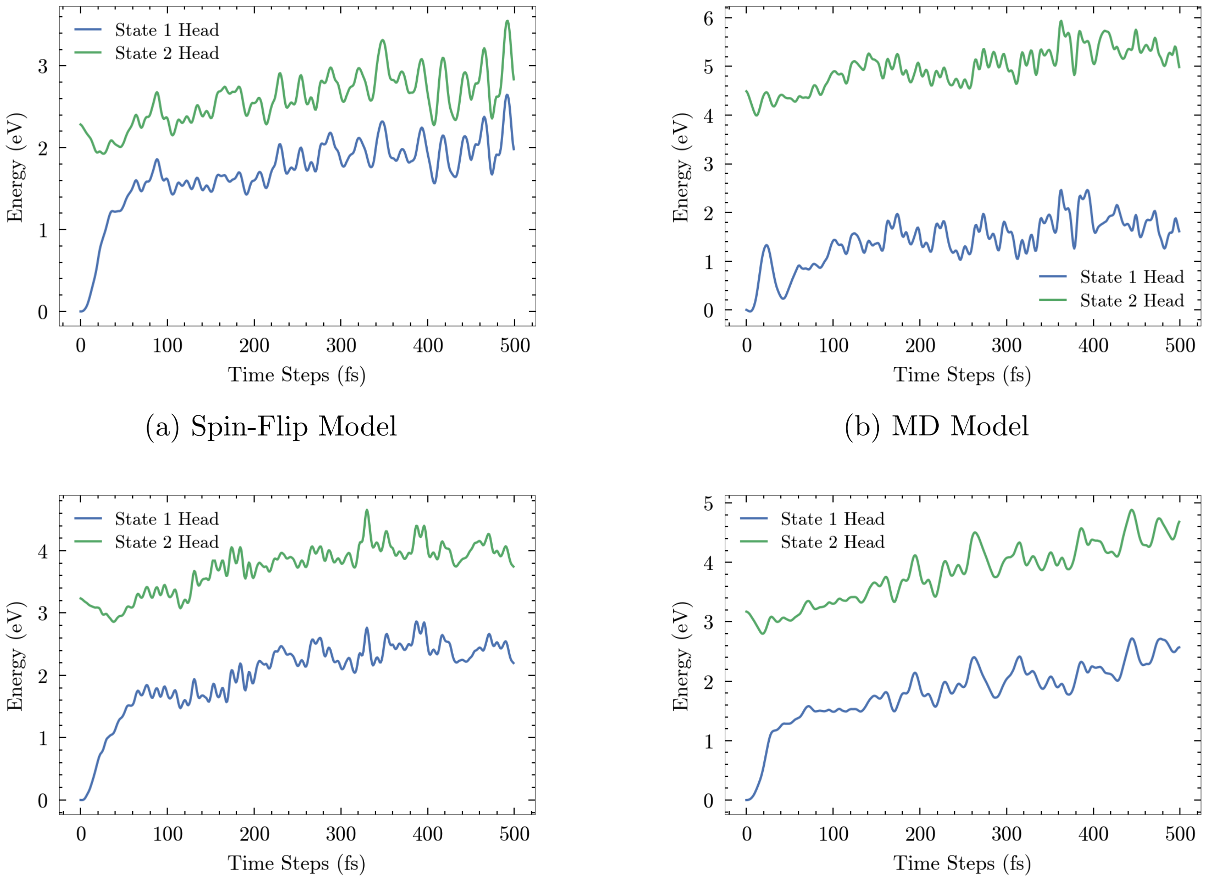
Furthermore, excitation to the conical intersection is expected to occur in less than 100fs [31], although not crossing paths, both, the *Spin-Flip Model* figure 7a and the *Combined Model* figure 7c show an accurate approximation of this excitation.

Further research would be required to determine if with the correct S_1 PES training data, a fine-tuned model can correctly approximate the conical intersection.

A..3 Molecular Dynamics benchmark

As mentioned in section III.D. 125 geometries were obtained by performing 25 simulations and taking 5 snapshot per simulation. The ground truth energy for each of the geometries was obtained via Spin-Flip Linear Response TDDFT. The RMSE of each model head was calculated and is shown in figure 8.

The figure illustrates that the model struggles to accurately learn the energy of high-



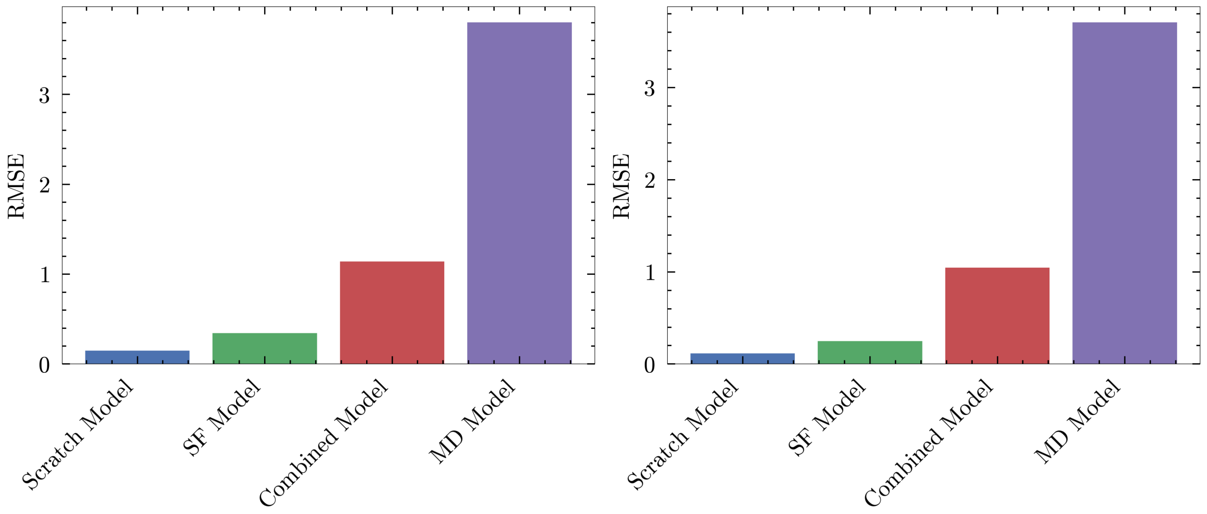
(a) Spin-Flip Model

(b) MD Model

(c) Combined Model

(d) Scratch Model

Figure 7: Thymine Dimer Molecular Dynamics simulations on each of the models. The *Scratch Model* refers to the model trained from scratch only on relaxed geometries. The *SF Model* is the fine-tuned model trained exclusively on relaxed geometries. The *Combined Model* is the fine-tuned model trained on both relaxed and molecular dynamics (MD) geometries, while the *MD Model* is trained solely on MD geometries.



(a) State 1

(b) State 2

Figure 8: Histogram of Energy RMSE for each of the four models on both Heads with Ground Truth: Spin-Flip TDDFT.

energy configurations. The error on the combined model is significant compared to the model trained solely on relaxed geometries. This suggests that while incorporating molecular dynamics geometries improves performance, further refinement or additional training strategies may be necessary to enhance accuracy for high-energy configurations.

However, the figure 8b should be approached with caution as the relaxed geometries dataset include spin contaminated configurations. Consequently, the fine-tuning process may face challenges in accurately distinguishing between different potential energy surfaces, as it is forced to treat them as equivalent.

B. Oxirane

B..1 Excited State Predictions

As with thymine dimer, both the Scratch and fine-tuned models are compared in their ability to predict the potential energy surface (PES) of oxirane. Figure 9 illustrates the differences between their predictions. Given the simplicity of the molecule, both models demonstrate high accuracy in capturing the PES of both electronic states. However the fine-tuned model greatly benefits from its previous training. This advantage is specially pronounced around the bond C-O bond breaking at 108°, greatly predicting the steep decline in potential energy of the ground state.

B..2 Multihead Molecular Dynamics

Oxirane trained models exhibit a much different behaviour than thymine dimer when predicting molecular dynamics. From figure 10 it can be seen that the fine-tuned models are able to simulate degenerate states. However none of them managed to exhibit the internal conversion when the State 1 energy is higher than the State 2. Furthermore, the excitation of oxirane on the fine-tuned models happen at timescale around 50 fs which is found to be in agreement with [32]. These findings highlight the importance of transfer learning techniques in quantum chemistry machine learning models, particularly when dealing with complex systems such as oxirane. The ability of fine-tuned models to accurately predict conical intersections suggests that they could be useful in modelling ultrafast photochemical reactions in large non-isolated systems, where traditional methods are too expensive to compute.

This also raises the question of whether GNNs and MLIPs can learn the topology associated with conical intersections and internal conversion processes, as opposed to the

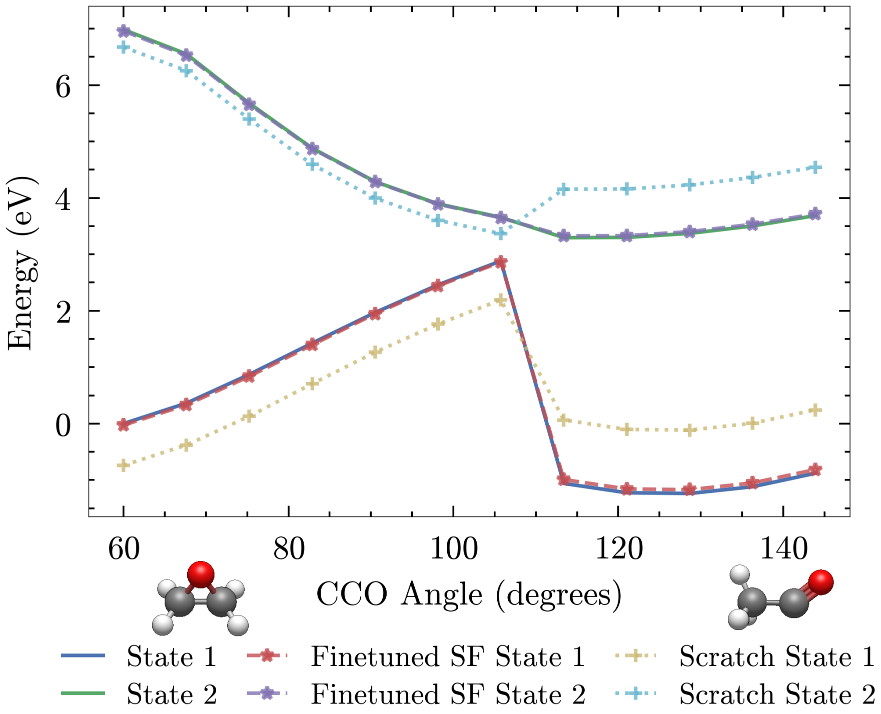


Figure 9: Scratch MACE-OFF vs Fine-tuned MACE-OFF predictions on the C-C-O angle opening of Oxirane. Fine-tuned model describes both surfaces with higher accuracy than the from scratch model.

intersystem crossings observed in this MD simulations.

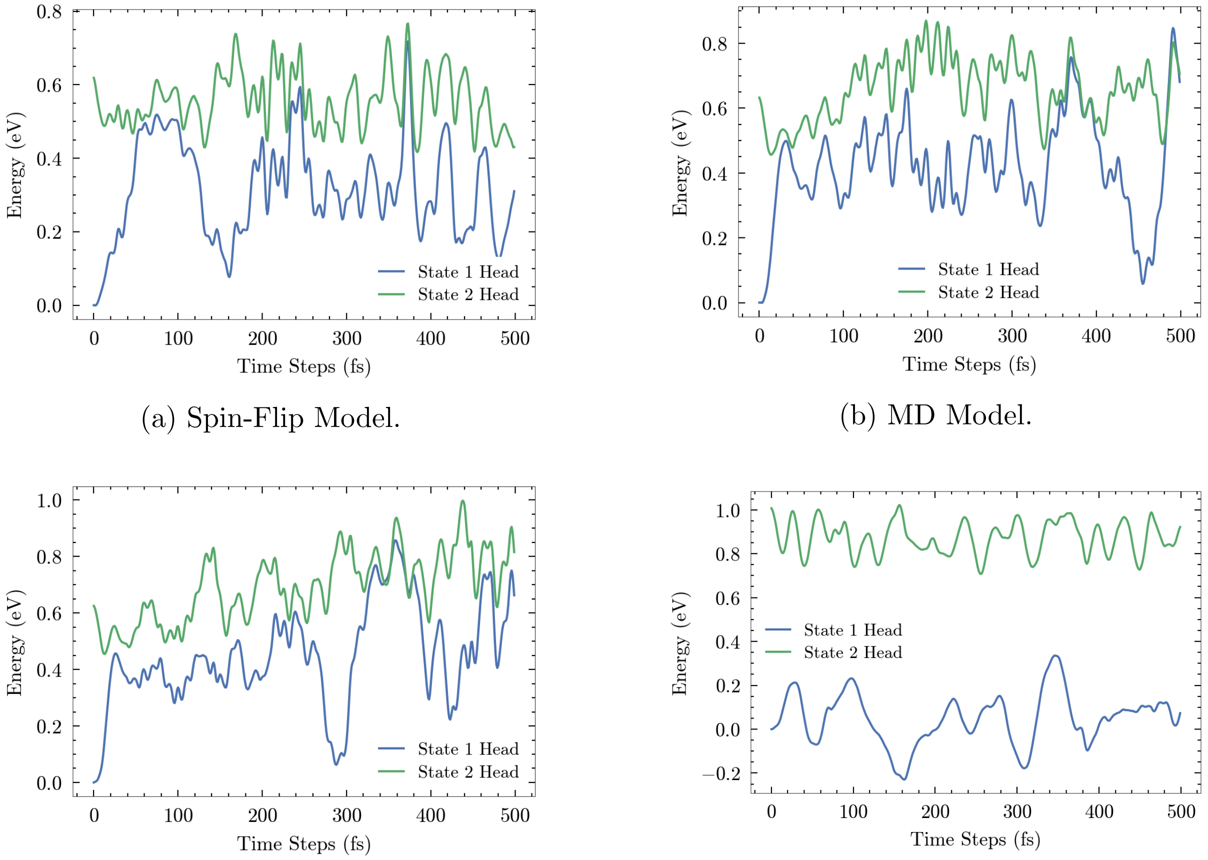
B..3 Molecular Dynamics Benchmark

The four models trained on oxirane were compared against 125 geometries sampled from 25 distinct molecular dynamics simulations. The histograms of this benchmark are shown in figure 11. Notably, the scratch model performs poorly on the excited state head, whereas the other models, benefiting from fine-tuning, demonstrate significantly improved performance. The fine-tuning allows these models to better capture the nuances of the excited state, leading to a more accurate representation of the system's behaviour.

V. Conclusion

This study demonstrates that fine-tuning foundational models is an effective strategy for obtaining a model with accuracy comparable to that of models trained from scratch for molecular dynamics, but with the advantage of requiring a significantly smaller dataset.

It was shown that, even with a relatively small dataset tailored to a specific molecular system, fine-tuned foundational models not only achieved accuracy comparable to models trained from scratch but, in some cases, exceeded their performance. Specifically, the



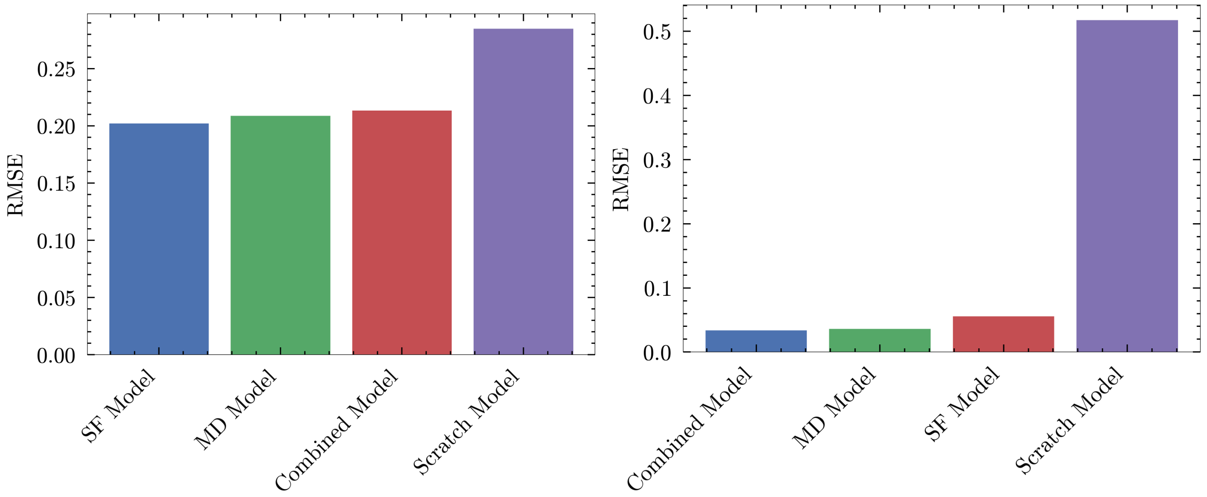
(a) Spin-Flip Model.

(b) MD Model.

(c) Combined Model.

(d) Scratch Model.

Figure 10: Oxirane Molecular Dynamics Simulations on each of the models. Given the reactivity of oxirane, energy crossings are expected, but the raw model cannot properly handle these crossings, showing the usefulness of fine-tuning over scratch models.



(a) State 1

(b) State 2

Figure 11: Histogram of energy RMSE for each of the four Oxirane models across both output heads, with Spin-Flip TDDFT as the ground truth. While all models perform relatively well in predicting the ground state (State 1), the Scratch Model exhibits significantly higher errors in the first excited state (State 2), indicating poor generalization to excited-state dynamics.

fine-tuned models demonstrated superior accuracy in predicting potential energy surfaces (PES), capturing intricate energetic variations more effectively.

Additionally, in energy predictions and molecular dynamics simulations, the fine-tuned models exhibited improved stability and consistency, suggesting that the pre-trained representations provided a more robust starting point for learning system-specific behaviours. These findings highlight the potential of transfer learning approaches in computational molecular modelling, reducing the dependence on large datasets while maintaining or even enhancing predictive performance.

However, without sufficiently accurate levels of theory to predict excited-state properties without compromising realism it is unreasonable to expect an MLIP to predict internal conversions or conical intersections. As in the case of Spin-Flip, with spin contamination issues. The challenge lies in generating high-quality data for these geometries.

Additionally, a question arises as to whether common training techniques can effectively train a multi-head model where, in the majority of its data, the first excited state is at a higher energy than the ground state, but in certain scenarios such as conical intersections this is not the case. Hence by choosing a head for each state, it is challenging to train such a two-head model to predict configurations where the ground state (head 1) is of higher energy than the first excited state (head 2). One potential approach could be to use two separate models: one for the ground state energy and another for the excitation energies. This strategy might improve the prediction of cases where excitation energies are close to zero or even negative.

Moreover, transfer learning may prove to be especially valuable when simulating large systems. With a small amount of data from an isolated reference system, the pre-trained knowledge can be effectively leveraged to account for long-range interactions when simulating the reference system within a reactant system.

In conclusion, while fine-tuning foundational models offers a promising approach to improving model accuracy with smaller datasets, the challenges related to the prediction of excited-state properties, internal conversions, conical intersections, and correct state following remain significant. Achieving reliable results in these areas requires high-quality data and advanced training strategies.

The potential for using multi-head models and transfer learning presents an efficient and accurate approach in the context of large systems and complex interactions. Future research is needed to refine these techniques and address the limitations posed by spin contamination and conical intersection topology.

References

[1] D. P. Kovács *et al.*, Mace-off: Transferable short range machine learning force fields for organic molecules, 2023.

[2] H. Li *et al.*, The Journal of Physical Chemistry Letters **13**, 5881–5893 (2022).

[3] A. Patra, A. I. Krylov, and S. Mallikarjun Sharada, The Journal of Chemical Physics **159** (2023).

[4] M. Born and R. Oppenheimer, Annalen der Physik **389**, 457–484 (1927).

[5] W. Domcke, D. R. Yarkony, and H. Köppel, *Conical Intersections: Theory, Computation and Experiment* (WORLD SCIENTIFIC, 2011).

[6] A. A. Beckstead, Y. Zhang, M. S. de Vries, and B. Kohler, Physical Chemistry Chemical Physics **18**, 24228–24238 (2016).

[7] D. S. Sholl and J. A. Steckel, *Density Functional Theory: A Practical Introduction* (Wiley, 2009).

[8] E. Engel and R. M. Dreizler, *Density Functional Theory: An Advanced Course* (Springer Berlin Heidelberg, 2011).

[9] R. Jones, Reviews of Modern Physics **87**, 897–923 (2015).

[10] P. Hohenberg and W. Kohn, Physical Review **136**, B864–B871 (1964).

[11] W. Kohn and L. J. Sham, Physical Review **140**, A1133–A1138 (1965).

[12] A. Najibi and L. Goerigk, Journal of Chemical Theory and Computation **14**, 5725–5738 (2018).

[13] A. Hellweg and D. Rappoport, Physical Chemistry Chemical Physics **17**, 1010–1017 (2015).

[14] *Fundamentals of Time-Dependent Density Functional Theory* (Springer Berlin Heidelberg, 2012).

[15] M. Sprengel, G. Ciaramella, and A. Borzì, SIAM Journal on Mathematical Analysis **49**, 1681–1704 (2017).

[16] D. Casanova and A. I. Krylov, Physical Chemistry Chemical Physics **22**, 4326–4342 (2020).

[17] W. Park, K. Komarov, S. Lee, and C. H. Choi, The Journal of Physical Chemistry Letters **14**, 8896–8908 (2023).

[18] J. Herbert and A. Mandal, (2022).

[19] P. Atkins and R. Friedman, *Molecular Quantum Mechanics* (Oxford University Press, 2010).

[20] D. Goldfarb, Mathematics of Computation **24**, 23–26 (1970).

[21] T. B. Blank, S. D. Brown, A. W. Calhoun, and D. J. Doren, The Journal of Chemical Physics **103**, 4129–4137 (1995).

[22] J. Behler and M. Parrinello, Physical Review Letters **98** (2007).

[23] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021.

[24] J. Sun, C. Yang, X. Ji, Q. Huang, and S. Wang, Towards dynamic message passing on graphs, 2024.

[25] F. Neese, WIREs Computational Molecular Science **2**, 73–78 (2011).

[26] J.-H. Li, T. J. Zuehlsdorff, M. C. Payne, and N. D. M. Hine, The Journal of Physical Chemistry C **122**, 11633–11640 (2018).

[27] J.-H. Li, T. J. Zuehlsdorff, M. C. Payne, and N. D. M. Hine, Physical Chemistry Chemical Physics **17**, 12065–12079 (2015).

[28] H. Moore *et al.*, Research data supporting ”mace-off23”, 2024.

[29] J. L. Alonso *et al.*, (2012).

[30] G. Zechmann and M. Barbatti, The Journal of Physical Chemistry A **112**, 8273–8279 (2008).

[31] W. J. Schreier *et al.*, Science **315**, 625–629 (2007).

[32] M. Krenz, U. Gerstmann, and W. G. Schmidt, ACS Omega **5**, 24057–24063 (2020).